

This document defines plans for the epochs, a changing schedule ☺, and scope of MADCAT Phase 2 data. Table 1 lists the data plans for the main MADCAT phase 2 Go/No Go evaluation.

**Table 1: Data to support the MADCAT Phase 2 Go/No Go evaluation. Summary of MADCAT Phase 2 data sets includes one language, Arabic, and two genres newswire (NW) and web-text (WB).**

	Training Set	Development Set	GNG Evaluation Set
<b>Source</b>	GALE P1 & P4 parallel training data, split into 3 releases.  <i>LDC-to-BBN</i>	GALE P2 development data (LDC2007E16)  <i>NIST/LDC Selects Data LDC-to-BBN/NIST</i>	GALE P3, possibly P4, evaluation data.  <i>LDC-to-NIST-to-BBN</i>
<b>Release Date</b>	#1: <b>FEB-9</b>	<i>NIST/LDC selection:</i> <b>FEB-23</b>	<i>LDC-to-NIST:</i> <b>OCT-15</b>
	#2: <b>MAY-11</b>	<i>LDC-to-BBN:</i> <b>MAY- 11</b>	<i>NIST-to-BBN:</i> <b>NOV-2</b>
	#3: <b>JUL-9</b>		
<b>Scribe</b>	<ul style="list-style-type: none"> <li>No more than (5) scribes per passage</li> <li>strive for scribe variability</li> </ul>	(3) scribes per passage	(3) scribes per passage  Only (2) scribes per passage to be post-edited <sup>1</sup>
<b>Scribe distribution</b>	50% new scribes 50% old scribes  (50% is a minimum goal for new scribes, more is better)	50% new scribes 50% from training	50% new scribes 50% from training  (new scribes for EVAL are unseen in TRN and DEV)  NIST to recommend scribe to data assignments <sup>2</sup>
<b>Passages</b>	Target 4,000 unique passages  <i>~125 words per passage</i>	200 unique passages - 100 newswire docs - 100 web texts  <i>125-135 English words per passage</i>	40k words - 80 pages per genre - 2 scribes per page (edited)  <i>~125 words per passage</i>
<b>Total</b>	~20,000 manuscripts	600 manuscripts - 200 passages * 3 scribes	480 manuscripts produced 320 manuscripts post edited

Note that items in red require actions.

Note in the P1 pilot:

- Source count = ~ 48K source words
- Reference count = ~61K English words

<sup>1</sup> NIST to decide before hand how to select the two scribes.

<sup>2</sup> NIST to produce a proposal to be approved by LDC.

## **GNG P2 Development Data Selection**

1. NIST will identify GALE P2 development data to serve as pages by providing LDC with a file listing:  
PAGE\_ID -> DOC\_ID : Start\_Seg : End\_Seg (word\_count)
2. NIST will use reference token count to estimate page size (~125-135 English words)
3. NIST identified 100 pages per genre (no more than 2 pages per DOCUMENT)

## **Choosing 2 of the 3 Scribes to be Included in Post Editing**

1. METHOD TBD ...
2. NIST will select the passage/scribe pairs to be included in the post editing process before the evaluation.
3. Selection will be made such that the post edited evaluation data set meets the scribe distribution of overlap with training data.
4. Selection will be made to include equal contributions from many scribes (to the extent possible)
- 5.

## **Scribe to Data Assignment**

1. METHOD TBD ...
2. LDC to provide NIST with an estimate of the total number of scribes available for evaluation data, for both new scribes and scribes already seen in training.
3. NIST will produce a doc assignment table that identifies overlap. For example

Scribe TRN1: doc-1, doc-2, doc-3, doc-7, doc-8, doc-9

Scribe TRN2: doc-1, doc-2, doc-3, doc-10, doc-11, doc-12

Scribe NEW1: doc-10, doc-11, doc-12, doc-4, doc-5, doc-6

Scribe NEW2: doc-

Table 2 lists the **NIST STRAWMAN PROPOSAL FOR** data plans for using Harmony data to support a challenge track in MADCAT Phase 2.

**Table 2: Data to support the MADCAT Challenge test for phase 2.**

	Training Set	Development Set	Challenge Evaluation Set
Source	Harmony Database <i>LDC-to-BBN</i>	Harmony Database <i>NIST/LDC Selects Data</i> <i>LDC-to-BBN/NIST</i>	Harmony Database <i>LDC-to-NIST-to-BBN</i>
Release Date	#1h: MAY-22 (10 passages)	JUL-1	<i>LDC-to-NIST: OCT-15</i>
	#2h: JUL-1 (140 passages)		<i>NIST-to-BBN: NOV-2</i>
Scribe	Limited to the Harmony scribe	Limited to the Harmony scribe	Limited to the Harmony scribe
Scribe distribution	Not applicable	Not applicable	Not applicable
Passage	150 passages (single image passage)	25 passages (single image passage)	50 passages <i>Total word count unknown</i> (single image passage)
Selection	LDC to hand select see selection guidelines	see selection guidelines	see selection guidelines
Evaluation			10K to be Post Edited

**Selection Guidelines:**

- METHOD TBD ...
- Basic MADCAT appropriateness (handwritten, Arabic ...)
- Non-**super**-controversial content causing severe embarrassment to the program or, human translators.
- No more than 10% of the passages for each data set (training, development, evaluation) shall have fewer than 50 words.
- No more than 10% of the passages for each data set shall have more than 250 words.
- No passage shall have more than 400 words.
- Should avoid documents that will crash the system, this is not to be a stress test.
  - o Avoid overlapping zones.
  - o Avoid complex tables (scoring issue)

Table 3: MADCAT P2 Evaluation Events

Current List of MADCAT Related Evaluation Events	Current SCHEDULE (as of FEB-9)	Possible Changes (as of FEB-26)
Training release 1 ( <i>GNG</i> )	FEB-9	
Training release 2 ( <i>GNG</i> )	MAY-11	
Training release 1h ( <i>Harmony</i> )	MAY-22	
Training release 2h ( <i>Harmony</i> )	JUL-1	
Training release 3 ( <i>GNG</i> )	JUL-9	
Development set ( <i>GNG</i> )	MAY-11	
Development set ( <i>Harmony</i> )	JUL-1	
Scoring tool updates for main Go/No Go evaluation	MAY-11	
Scoring tool for Challenge evaluation	AUG-29	
P2 evaluation begins (BBN process data)	NOV-2	OCT-28
P2 evaluation ends (Results due at NIST)	Nov 20	NOV-18
NIST sends LDC Post Editing Kits	NOV-30	
Post Editing Completes	JAN-31-10	JAN-27-10
Final scores due at DARPA	FEB-3-10	JAN-29-10

Unresolved questions:

- 1) Can Harmony Challenge data be processed alongside GNG evaluation data? Is there a need to separate the processing?
- 2) How much Harmony data to be post edited?
- 3)